



HAL
open science

Scale and shift invariant time/frequency representation using auditory statistics: application to rhythm description

Ugo Marchand, Geoffroy Peeters

► **To cite this version:**

Ugo Marchand, Geoffroy Peeters. Scale and shift invariant time/frequency representation using auditory statistics: application to rhythm description . IEEE International Workshop on Machine Learning for Signal Processing, Sep 2016, Vietri Sul Mare, Italy. 2016. hal-01368888

HAL Id: hal-01368888

<https://hal.science/hal-01368888>

Submitted on 20 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Scale and Shift Invariant Time/Frequency Representation Using Auditory Statistics: Application to Rhythm Description.

Outline

Objectives:

- creating a representation of the audio signal that differentiate musical rhythms

Constraints:

- creating a representation which is invariant to tempo and to temporal-shifts

Propositions:

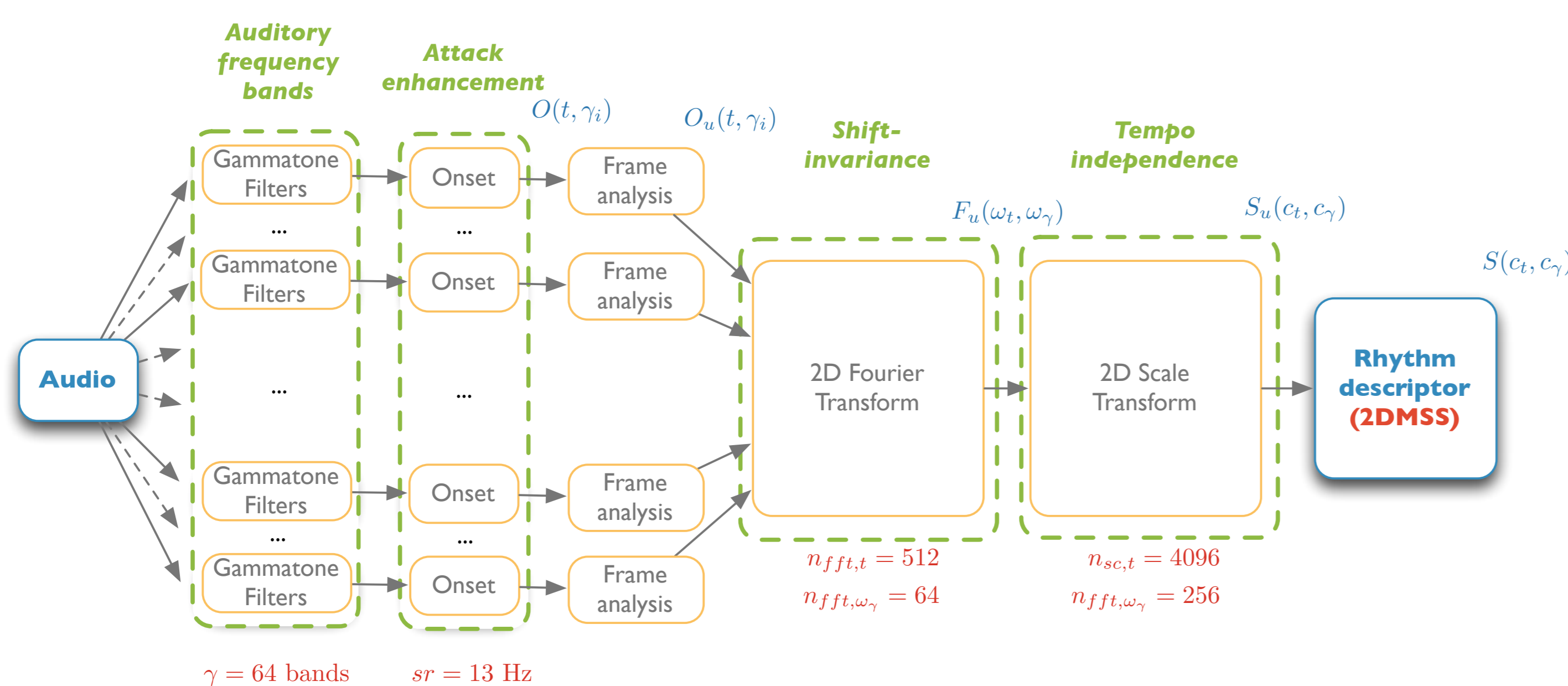
- Two new 2D (time/frequency) representations of the audio content: 2DMSS and MASSS
- New Dataset

Applications:

- use this representation to do auto-tagging, search by similarity

Rhythm description

Method 2DMSS



2DMSS= 2D Fourier Transform, followed by a 2D Scale Transform known as Fourier-Mellin Transform in Image processing
2D Scale Transform:

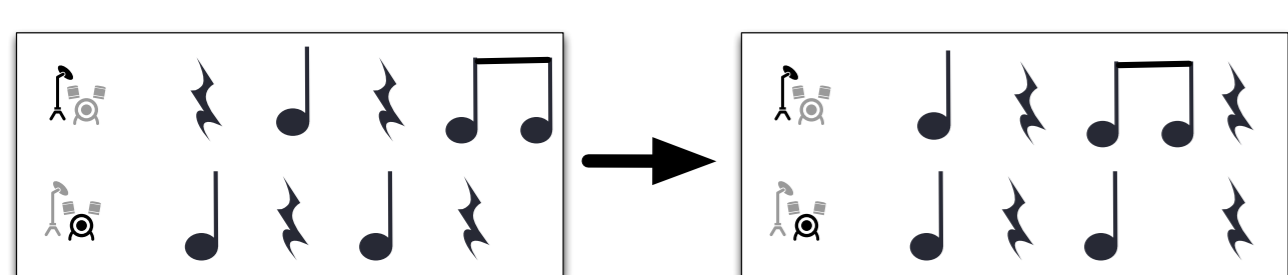
$$S(c_t, c_\omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(X(e^t, e^\omega) e^{\omega/2} e^{t/2} \right) e^{-jc_\omega \omega - jc_t t} d\omega dt$$

Pros/Cons

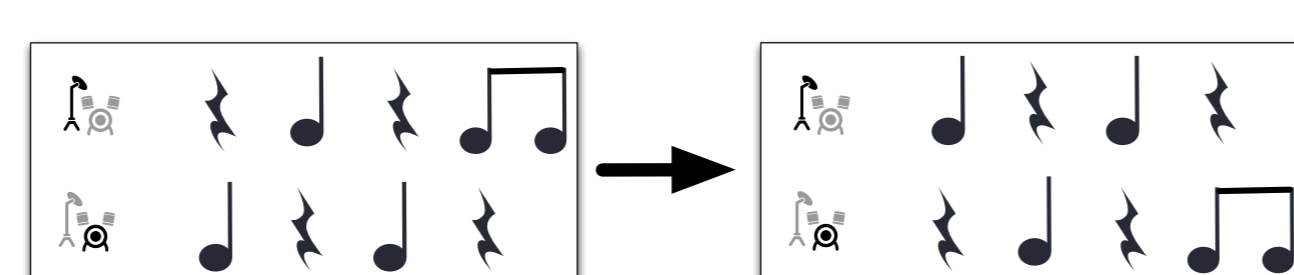
Pros: models the relationship between frequency bands and time bins with shift-invariance and scale-invariance

Cons: produces also shift-invariance over frequencies which is an undesired property

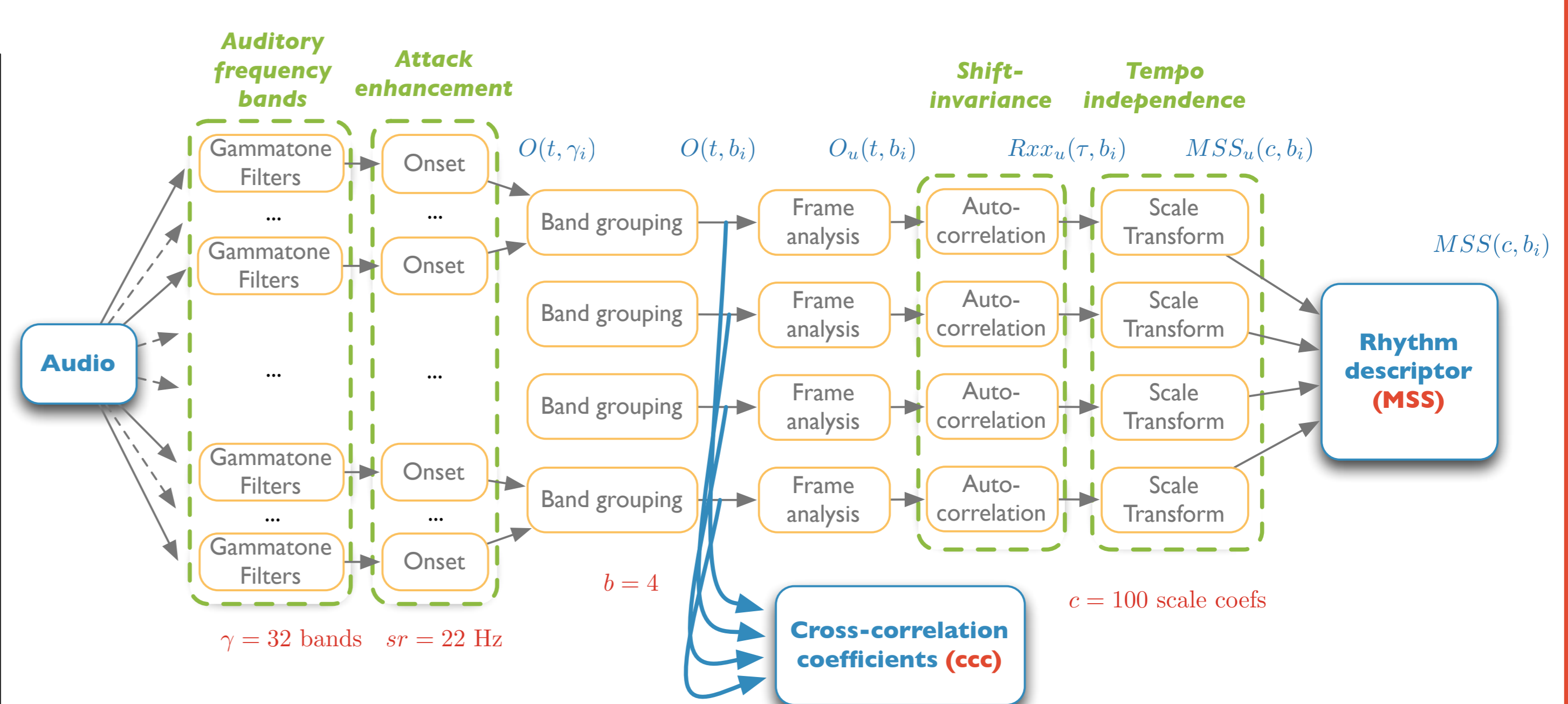
Can differentiate:



But not:



Method MASSS

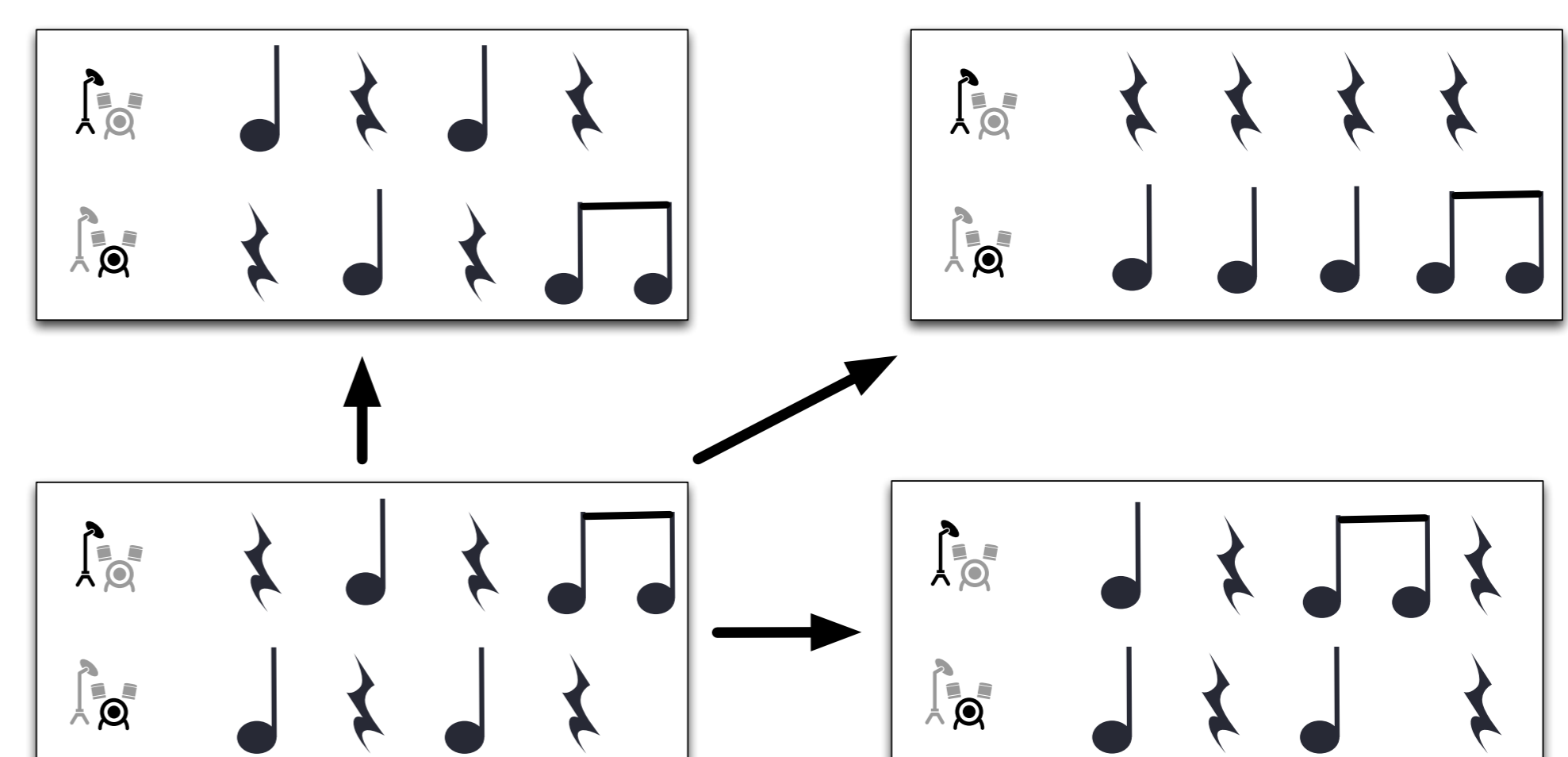


MASSS = Modulation Scale Spectrum (MSS) + Auditory Statistics (ccc) [McDermott, 2013]

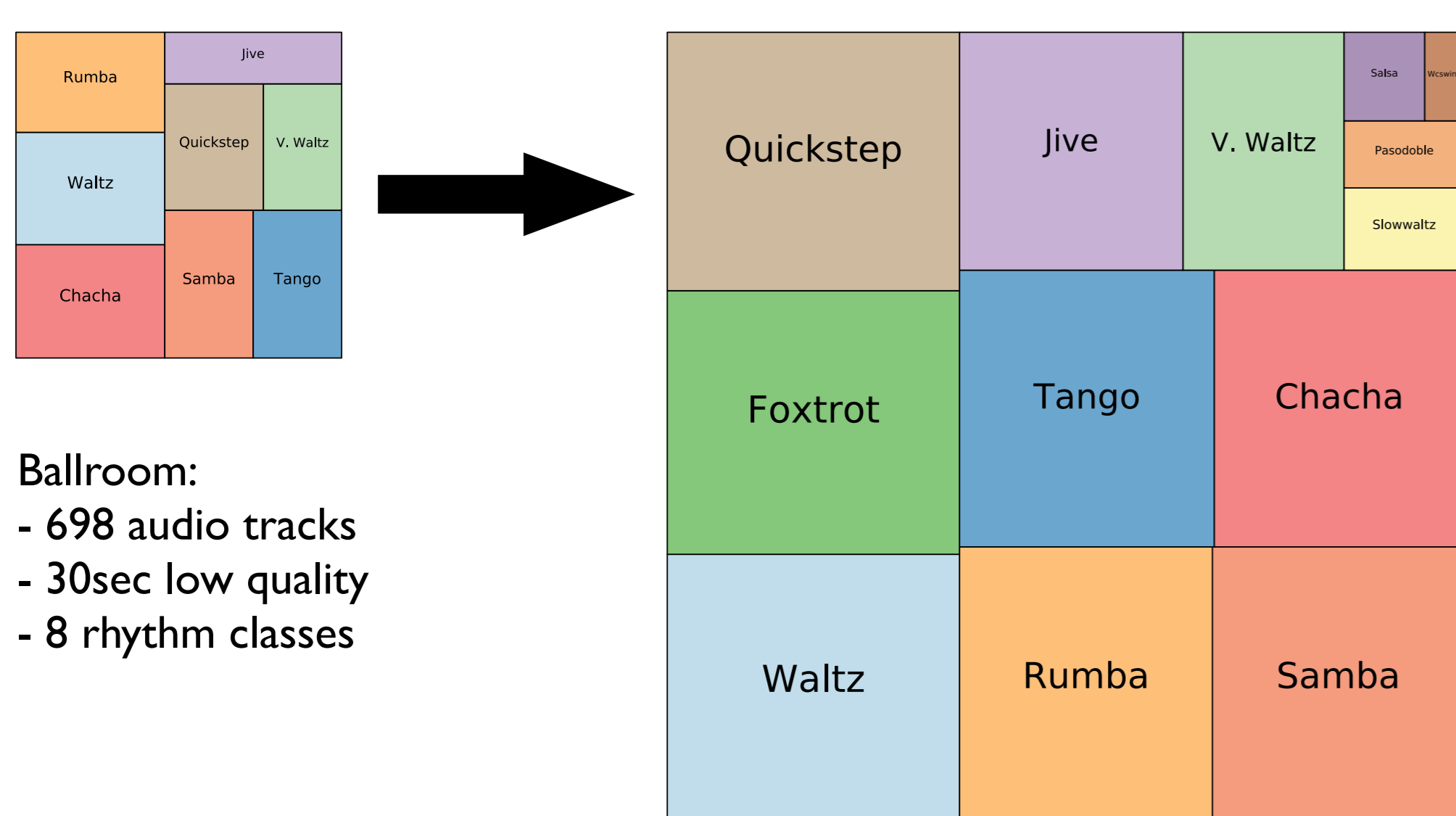
Auditory Statistics (ccc) = Cross-correlations between different auditory frequency bands

Late-fusion of MSS and ccc

Can differentiate:



New dataset : Extended Ballroom

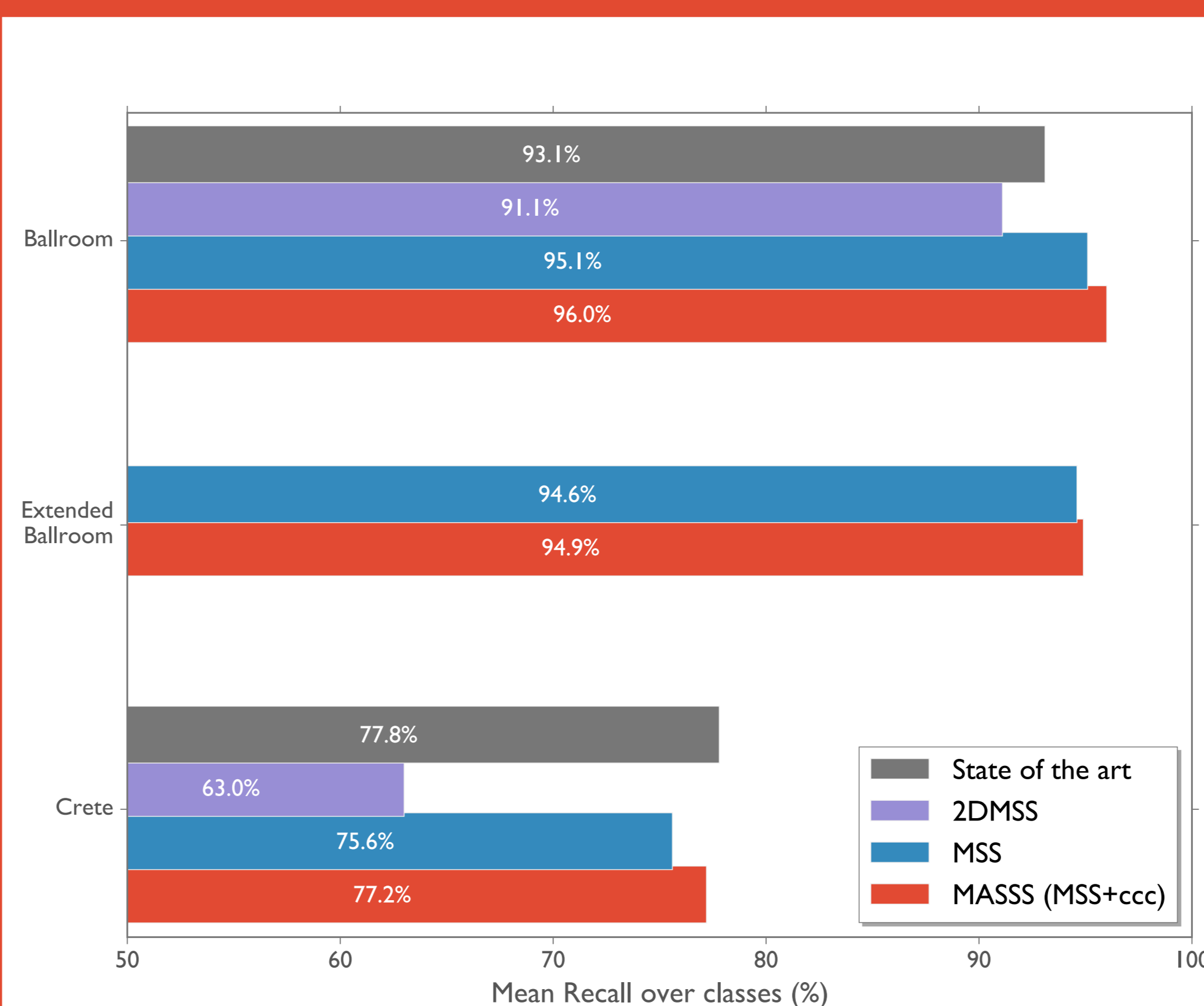


Ballroom:
- 698 audio tracks
- 30sec low quality
- 8 rhythm classes

Extended Ballroom:
- 4.180 audio tracks
- 30sec high-quality
- 9+4 rhythm classes
- similarity annotations



Results



Classification

- SVM (MSS, 2DMSS, ccc models)
- Logistic Regression (late-fusion)

Analysis of results:

- 2DMSS is not sufficient
- MASSS
- improves state-of-the-art method by 3% on Ballroom
- equals state-of-the-art on Cretan dances dataset.

Conclusion:

- modeling frequency bands inter-relationship through auditory statistics improves rhythm description